On an ε-Optimal Policy of Continuous Time Markov Decision Processes(MATHEMATICAL OPTIMIZATION AND ITS APPLICATIONS)

AUTHOR(S):

星野, 満博; 田中, 謙輔

数理解析研究所講究録
第835巻 1993年 81-88

On an ε-Optimal Policy of Continuous Time Markov Decision Processes

星野満博　（新潟大自然科）
田中謙輔　（新潟大理）

本報告では、一般の状態空間上の割引率 $\alpha > 0$ を伴う連続時間マルコフ決定過程を扱う。一般に、このような動的決定問題において最適政策が存在しないこともあり、この場合、ε-最適政策について論ずるのが自然であろう。最適政策、ε-最適政策に関しては [1] などでも論じられている。[1] では、ε-最適政策の存在性が示されているが、ここでは、これと異なり、ある ε-最適政策が存在するとき、損失関数に適当な修正を施すことにより、この ε-最適政策が、最適政策になるようにできることを議論している。

1 Time-homogeneous Markov process に関する contraction semigroup について

Zを complete separable metric space とし、β_Zを Z上の Borel σ-algebra とする。また、各 $z \in Z$に対して、$(\Omega, \mathcal{F}, P_z)$ を probability space 、$\{Z_t; t \geq 0\}$ を probability space $(\Omega, \mathcal{F}, P_z)$ をもつ Z上の time-homogeneous Markov process とする。

Def. 1.1　$\{Z_t; t \geq 0\}$ の transition probability function Pを、

$$P(z; t, \Gamma) = P_z\{Z_t \in \Gamma\} \qquad z \in Z, \Gamma \in \beta_Z, t \geq 0 \tag{1.1}$$

によって定義し、$B(Z)$ を Z上の有界かつβ_Z-可測な関数の集合とする。このとき、$B(Z)$ は、Banach space になる。 ただし、norm は、$\|f\| = \sup_{z \in Z} |f(z)|$ とする。

Def. 1.2　各 $t \geq 0$ に対して、operator $T_t : B(Z) \longrightarrow B(Z)$,$t \geq 0$ を、次のように定義する。

$$\begin{aligned} T_t f(z) &= E_z[f(Z_t)] \\ &= E[f(Z_t)|Z_0 = z] \\ &= \int_Z f(y) P(z; t, dy) \qquad f \in B(Z), z \in Z \end{aligned} \tag{1.2}$$

このとき、族 $\{T_t; t \geq 0\}$ は $B(Z)$ 上の bounded linear operator の contraction semigroup になる。

Def. 1.3　$\{f_n; n \geq 1\} \subset B(Z)$ に対して、

$$\operatorname*{wlim}_{n\to\infty} f_n = f \iff (i)\ \lim_{n\to\infty} f_n(z) = f(z) \quad \text{for each } z \in Z.$$
$$(ii)\ \{\|f_n\|; n \geq 1\} : \text{bounded sequence.}$$

$B_0 \equiv \{f \in B(Z) | \operatorname*{wlim}_{t\downarrow 0} T_t f = f\}$　と定義する。

Def. 1.4　contraction semigroup $\{T_t; t \geq 0\}$ の weak infinitesimal operator A を、

$$Af = \operatorname*{wlim}_{t\downarrow 0} \frac{T_t f - f}{t} \tag{1.3}$$

によって定義し、A の定義域を、$\mathcal{D}(A) \equiv \{f \in B_0 | \exists Af \in B_0\}$　とする。

2 Formulation

次の 6 個の対象物から成る dynamic decision model を考える。

(a) state space $\mathfrak{X}$: complete separable metric space の nonempty Borel subset.
$\beta_{\mathfrak{X}}$: $\mathfrak{X}$ の Borel σ-algebra.

(b) action space $\mathcal{A}$: complete separable metric space の nonempty Borel subset.
$\beta_{\mathcal{A}}$: $\mathcal{A}$ の Borel σ-algebra.

(c) time set $\mathcal{T} = [0, \infty)$, β_t : σ-algebra.

(d) low of motion.　ある time $t_0 \in \mathcal{T}$ において、action $a \in \mathcal{A}$ を選んだとき、process $\{X_t; t \geq 0\}$ の time $t_0 \in \mathcal{T}$ における確率的動きは、weak infinitesimal operator A_a と $x_{t_0} \in \mathfrak{X}$ によって定まる。

(e) loss function $r : [0, \infty) \times \mathfrak{X} \times \mathcal{A} \to R$, measurable.

(f) admissible policy の集合 D_Aは Markov policy から成る。

policy について

· π : (nonrandomized) Markov policy $\iff$ $\pi : [0, \infty) \times \mathfrak{X} \to \mathcal{A}$, $\beta_t \times \beta_{\mathfrak{X}}$-measurable.

· Markov policy が time に依存しない場合 stationary policy と呼ばれる。

$$i.e. \quad \pi(t, x) = \pi(0, x) = \pi(x) \qquad \forall t \geq 0, x \in \mathfrak{X}$$

前提　各 $\pi \in D_A$ に対して、次の性質を持つ、stochastic process $\{X_t; t \geq 0\}$ が存在するものとする。

(i)　$\{X_t; t \geq 0\}$: strongly measurable , strong Markov process.

(ii)　任意の time t_0 における Markov process $\{X_t; t \geq 0\}$ の確率的動きは、weak infinitesimal operator $A_{\pi(t_0, x_{t_0})}$ によって決まる。

(iii)　$\{X_t; t \geq 0\}$ の殆どすべての sample path は、右連続かつ左側極限を持ち、任意の有限時間区間において、不連続点は高々有限個である。

policy πを用いたときの $\{X_t; t \geq 0\}$ の transition probability function P_π を

$$P_\pi(s, x; s + t, \Gamma) = P_\pi\{X_{s+t} \in \Gamma | X_s = x\} \qquad s \geq 0, t \geq 0, x \in \mathfrak{X}, \Gamma \in \beta_{\mathfrak{X}} \tag{2.1}$$

とする。policy πによって導かれる Markov process $\{X_t; t \geq 0\}$ は、time-homogeneous であるとは限らない。しかし、state space を拡張することにより各 $\pi \in D_A$ に対して、time-homogeneous Markov process を得ることができる。

直積空間 $([0,\infty), \beta_t) \times (\mathfrak{X}, \beta_{\mathfrak{X}})$ 上で定義された、各 $\pi \in D_A$ に対する 2 変数 process $\{(t, X_t); t \geq 0\}$ の transition probability function H_π を次のように定義する。

(2.2)
$$
\begin{aligned}
& H_\pi(t_1, x; t_2, J, \Gamma) \\
& = P_\pi\{(t_1 + t_2, X_{t_1+t_2}) \in J \times \Gamma | (t, X_t) = (t_1, x)\} \\
& = \begin{cases} P_\pi(t_1, x; t_1 + t_2, \Gamma) & \text{if } J \cap [t_1 + t_2, \infty) \neq \phi \\ 0 & \text{if } J \cap [t_1 + t_2, \infty) = \phi \end{cases}
\end{aligned}
$$

$$t_1 \geq 0\ ,\ t_2 \geq 0\ ,\ J \in \beta_t$$

このとき、$\{(t, X_t); t \geq 0\}$ は time-homogeneous Markov process となる。$Z_t = (t, X_t)$ とおくと、第 1 章のように各 $\pi \in D_A$ に対して $\{(t, X_t); t \geq 0\}$ の contraction semigroup を $\{T_t^\pi; t \geq 0\}$ とし、その weak infinitesimal operator を A_π と表すことにして、B_0^π 、$\mathcal{D}(A_\pi)$ を同様に定義することができる。

本報告における continuous time Markov decision process において次のように仮定する。

条件 A

(a) $\exists B_0 \in B(Z)\ s.t.\ B_0 \neq \phi\ ,\ B_0 \subset \bigcap_{\pi \in D_A} B_0^\pi$

(b) $\exists \mathcal{D}(A) \subset B(Z)\ s.t.\ \mathcal{D}(A) \neq \phi\ ,\ \mathcal{D}(A) \subset \bigcap_{\pi \in D_A} \mathcal{D}(A_\pi)$

(c) $T_t^\pi 1 = 1\ ,\ A_\pi 1 = 0\ ,\ \pi \in D_A$

(d) $\exists M < \infty$ s.t. $|r(t, x, a)| \leq M\ ,\quad t \geq 0, x \in \mathfrak{X}, a \in \mathcal{A}$

(e) 各$\pi \in D_A$に対して、$r_\pi : Z \to R$ を $r_\pi(t, x) = r(t, x, \pi(t, x))$ によって定義するとき、$r_\pi \in B_0$

$\alpha > 0$ を割引率として、次のような total expected discounted loss $V_\pi(t, x)$ を考える。

(2.3)
$$
\begin{aligned}
V_\pi(t, x) &\equiv E_\pi\left[\int_t^\infty e^{-\alpha(\tau - t)} r(\tau, X_\tau, \pi(\tau, X_\tau)) d\tau | X_t = x\right] \\
&= \int_t^\infty e^{-\alpha(\tau - t)} E_\pi[r(\tau, X_\tau, \pi(\tau, X_\tau)) | X_t = x] d\tau \\
&= \int_0^\infty e^{-\alpha\tau} T_\tau^\pi r_\pi(t, x) d\tau \qquad t \geq 0\ ,\ x \in \mathfrak{X} \quad ([1]\text{参照})
\end{aligned}
$$

これは、time $t \geq 0$ において、state $x \in \mathfrak{X}$ が観測されたという初期状態からスタートして、policy $\pi \in D_A$ を用いた場合の total expected discounted loss を表している。

Def.2.1 optimal discounted loss function $V_* : [0, \infty) \times \mathfrak{X} \to R$ を、

$$V_*(t, x) = \inf_{\pi \in D_A} V_\pi(t, x) \qquad t \geq 0, x \in \mathfrak{X}$$

によって定義する。

Def.2.2 $\pi^* \in D_A$: optimal in $D_A \iff V_{\pi^*}(t,x) = V_*(t,x) \qquad \forall t \geq 0, x \in \mathfrak{X}$

Def.2.3 各 $\varepsilon > 0$ に対して、

$\pi^* \in D_A$: ε-optimal in $D_A \iff V_{\pi^*}(t,x) \leq V_*(t,x) + \varepsilon \qquad \forall t \geq 0, x \in \mathfrak{X}$

3 Main Results

3.1 some conditions

我々のアプローチは次の Ekeland の定理に基づている。

Ekeland の定理

X : complete metric space , d : metric.

F : X $\to (-\infty, \infty]$, $F \not\equiv \infty$, 下に有界, 下半連続.

と仮定する。このとき、ある $\varepsilon > 0$ に対して $F(u) \leq \inf_{x \in X} F(x) + \varepsilon$ となる $u \in X$ について、次の 3 条件を満たす $v \in X$ が存在する。

(3.1) $$F(v) \leq F(u)$$

(3.2) $$d(u,v) \leq 1$$

(3.3) vと異なる全ての $w \in X$に対して、 $F(w) > F(v) - \varepsilon d(v,w)$

Ekeland の定理を我々の動的決定問題に適用するために、次の 2 つの条件を仮定する。

条件 B

(a) $\mathcal{A}$: complete metric space. ρ : $\mathcal{A}$ の metric.

(b) 与えられた $\bar{\pi} \in D_A$ に対して、

- $V_{\bar{\pi}} \in \mathcal{D}(A_a)$, $a \in \mathcal{A}$
- $V_{\bar{\pi}} \in \mathcal{D}(A)$
- $\rho(\pi_1(\cdot,\cdot), \pi_2(\cdot,\cdot)) \in B_0, \ \forall \pi_1, \pi_2 \in D_A$

(c) 与えられた $t_0 \geq 0$,$x_0 \in \mathfrak{X}$,$\bar{\pi} \in D_A$ に対して、$F_\cdot(t_0, x_0) : \mathcal{A} \to R$ を

$$F_a(t_0,x_0) = r(t_0,x_0,a) + A_a V_{\bar{\pi}}(t_0,x_0) \qquad , a \in \mathcal{A}$$

によって定義するとき、$F_\cdot(t_0,x_0)$ は、$\mathcal{A}$ 上、下に有界かつ下半連続である。

条件 B を仮定するとき、Ekeland の定理により、各 $\varepsilon > 0$,$t_0 \geq 0$,$x_0 \in \mathfrak{X}$,$\bar{\pi} \in D_A$ に対して、

$\exists a^* \in \mathcal{A}$ $s.t.$

(3.4) $a \neq a^*$ である全ての $a \in \mathcal{A}$ に対して、

$$r(t_0,x_0,a) + A_a V_{\bar{\pi}}(t_0,x_0) > r(t_0,x_0,a^*) + A_{a^*} V_{\bar{\pi}}(t_0,x_0) - \alpha\varepsilon\rho(a^*,a)$$

条件 C (3.4) の $a^* \in \mathcal{A}$ に関して、

(a) $D_A^*(t_0,x_0) \equiv \{a^* \in \mathcal{A} \mid a^* \neq \overline{\pi}(t_0,x_0)$,(3.4) をみたす $\} \neq \phi$
(b) $\exists \pi^* \in D_A$ $s.t.$ $\pi^*(t_0,x_0) \in D_A^*(t_0,x_0)$

条件 B,C を仮定するとき、各 $\varepsilon > 0, t_0 \geq 0, x_0 \in \mathfrak{X}, \overline{\pi} \in D_A$ に対して、(3.4) より $a \neq \pi^*(t_0,x_0)$ である全ての $a \in \mathcal{A}$ に対して、

$$r(t_0,x_0,a) + A_a V_{\overline{\pi}}(t_0,x_0) > r_{\pi^*}(t_0,x_0) + A_{\pi^*} V_{\overline{\pi}}(t_0,x_0) - \alpha\varepsilon\rho(\pi^*(t_0,x_0),a)$$

特に、a として、$\overline{\pi}(t_0,x_0) \in \mathcal{A}$ をとると、

$$r_{\overline{\pi}}(t_0,x_0) + A_{\overline{\pi}} V_{\overline{\pi}}(t_0,x_0) > r_{\pi^*}(t_0,x_0) + A_{\pi^*} V_{\overline{\pi}}(t_0,x_0) - \alpha\varepsilon\rho(\pi^*(t_0,x_0),\overline{\pi}(t_0,x_0))$$

[1 , theorem4.1] により、

(3.5) $$\alpha V_{\overline{\pi}}(t_0,x_0) > r_{\pi^*}(t_0,x_0) + A_{\pi^*} V_{\overline{\pi}}(t_0,x_0) - \alpha\varepsilon\rho(\pi^*(t_0,x_0),\overline{\pi}(t_0,x_0))$$

よって、

(3.6) $\exists \delta > 0$ $s.t.$

$$\alpha V_{\overline{\pi}}(t_0,x_0) > r_{\pi^*}(t_0,x_0) + A_{\pi^*} V_{\overline{\pi}}(t_0,x_0) - \alpha\varepsilon\rho(\pi^*(t_0,x_0),\overline{\pi}(t_0,x_0)) + \delta$$

更に、次の条件を仮定する。

条件 D $\exists \tau_0 > 0, \exists \delta_0 > 0$ $s.t.$ $P_{\pi^*}(t_0,x_0;t_0+\tau,I(x_0)) > 0$, $0 \leq \tau < \tau_0$
where $I(x_0) \equiv \{x \in \mathfrak{X} \mid d(x_0,x) \leq \delta_0\}$, $d : \mathfrak{X}$ の metric.

$S(t_0,x_0) \equiv \{(t,x) \mid t_0 \leq t < t_0 + \tau_0,\ d(x_0,x) \leq \delta_0\}$ とする。

条件 E (3.6) が成り立つとき、

(3.7) $$\alpha V_{\overline{\pi}}(t,x) > r_{\pi^*}(t,x) + A_{\pi^*} V_{\overline{\pi}}(t,x) - \alpha\varepsilon\rho(\pi^*(t,x),\overline{\pi}(t,x)) + \delta \quad \text{if} \quad (t,x) \in S(t_0,x_0)$$

が成り立つ。

条件 F 与えられた、$\varepsilon > 0, t_0 \geq 0, x_0 \in \mathfrak{X}, \overline{\pi} \in D_A, \pi^* \in D_A$ に対して、π' を

$$\pi'(t,x) \equiv \begin{cases} \pi^*(t,x) & \text{if } (t,x) \in S(t_0,x_0) \\ \overline{\pi}(t,x) & \text{if } (t,x) \notin S(t_0,x_0) \end{cases}$$

によって定義するとき $\pi' \in D_A$ である。

定理 1 各 $\varepsilon > 0$,$t_0 \geq 0$,$x_0 \in \mathfrak{X}$,$\overline{\pi} \in D_A$ に対して、条件 A～ F を仮定するとき、$\exists \pi' \in D_A$ $s.t.$

(3.8) $$V_{\pi'}(t_0,x_0) - \alpha\varepsilon \int_0^{\tau_0} e^{-\alpha\tau} \int_{I(x_0)} \rho(\pi'(t_0+\tau,y),\overline{\pi}(t_0+\tau,y)) P_{\pi'}(t_0,x_0;t_0+\tau,dy)d\tau < V_{\overline{\pi}}(t_0,x_0)$$

証明 $\varepsilon > 0$,$t_0 \geq 0$,$x_0 \in \mathfrak{X}$,$\overline{\pi} \in D_A$ を、固定する。(3.7) により $0 \leq \tau < \tau_0$, $d(x_0, x) \geq \delta_0$ のとき、

$$\alpha V_{\overline{\pi}}(t_0+\tau, x) > r_{\pi^*}(t_0+\tau, x) + A_{\pi^*} V_{\overline{\pi}}(t_0+\tau, x) - \alpha\varepsilon\rho(\pi^*(t_0+\tau, x), \overline{\pi}(t_0+\tau, x)) + \delta$$

$$\therefore \quad \int_{I(x_0)} \alpha V_{\overline{\pi}}(t_0+\tau, y) P_{\pi'}(t_0, x_0; t_0+\tau, dy)$$
$$\geq \int_{I(x_0)} \{r_{\pi^*}(t_0+\tau, y) + A_{\pi^*} V_{\overline{\pi}}(t_0+\tau, y)$$
$$- \alpha\varepsilon\rho(\pi^*(t_0+\tau, y), \overline{\pi}(t_0+\tau, y)) + \delta\} P_{\pi'}(t_0, x_0; t_0+\tau, dy)$$
$$0 \leq \tau < \tau_0$$

π' の定義より

$$\int_{I(x_0)} \alpha V_{\overline{\pi}}(t_0+\tau, y) P_{\pi'}(t_0, x_0; t_0+\tau, dy) \tag{3.9}$$
$$\geq \int_{I(x_0)} \{(r_{\pi'} + A_{\pi'} V_{\overline{\pi}})(t_0+\tau, y)$$
$$- \alpha\varepsilon\rho(\pi^*(t_0+\tau, y), \overline{\pi}(t_0+\tau, y)) + \delta\} P_{\pi'}(t_0, x_0; t_0+\tau, dy)$$
$$0 \leq \tau < \tau_0$$

一方、$\alpha V_{\overline{\pi}} = r_{\overline{\pi}} + A_{\overline{\pi}} V_{\overline{\pi}}$ と π' の定義により、

$$\int_{I(x_0)^c} \alpha V_{\overline{\pi}}(t_0+\tau, y) P_{\pi'}(t_0, x_0; t_0+\tau, dy) \tag{3.10}$$
$$= \int_{I(x_0)^c} (r_{\pi'} + A_{\pi'} V_{\overline{\pi}})(t_0+\tau, y) P_{\pi'}(t_0, x_0; t_0+\tau, dy) \qquad \tau \geq 0$$

(3.9) ,(3.10) を加えると、次の不等式が得られる。

$$T_\tau^{\pi'} \alpha V_{\overline{\pi}}(t_0, x_0) \geq T_\tau^{\pi'} (r_{\pi'} + A_{\pi'} V_{\overline{\pi}})(t_0, x_0)$$
$$- \int_{I(x_0)} \{\alpha\varepsilon\rho(\pi'(t_0+\tau, y), \overline{\pi}(t_0+\tau, y)) - \delta\} P_{\pi'}(t_0, x_0; t_0+\tau, dy)$$
$$0 \leq \tau < \tau_0$$

よって

$$-\frac{d^+}{d\tau}(e^{-\alpha\tau} T_\tau^{\pi'} V_{\overline{\pi}}(t_0, x_0)) \tag{3.11}$$
$$\geq e^{-\alpha\tau} T_\tau^{\pi'} r_{\pi'}(t_0, x_0)$$
$$- e^{-\alpha\tau} \int_{I(x_0)} \{\alpha\varepsilon\rho(\pi'(t_0+\tau, y), \overline{\pi}(t_0+\tau, y)) - \delta\} P_{\pi'}(t_0, x_0; t_0+\tau, dy)$$
$$0 \leq \tau < \tau_0$$

また、$\alpha V_{\overline{\pi}} = r_{\overline{\pi}} + A_{\overline{\pi}} V_{\overline{\pi}}$ より、(3.11) のときと同様に次の等式が得られる。

$$-\frac{d^+}{d\tau}(e^{-\alpha\tau} T_\tau^{\pi'} V_{\overline{\pi}}(t_0, x_0)) = e^{-\alpha\tau} T_\tau^{\pi'} V_{\pi'}(t_0, x_0) \qquad \tau \geq \tau_0 \tag{3.12}$$

(3.11) , (3.12) をそれぞれ $0 \leq \tau < \tau_0$, $\tau_0 \leq \tau < \infty$ の範囲で積分して加えると、(3.8) が得られる。 ■

3.2 条件 C(a) が成り立たない場合について

定理 1 を導く仮定において、条件 C(a) を仮定したが、条件 C(a) が成り立たない場合も、次のような結果が得られる。

定理 2 ある $\bar{\pi} \in D_A$ に対して、

$$\begin{cases} \text{条件 } A\,,\ B(a)\,,\ B(b) \\ \text{条件 } B(c) \quad \text{for each t}\geq 0\,,\ x \in \mathfrak{X} \\ \text{条件 } C' \qquad D_A^*(t,x) = \phi\,,\ \forall t \geq 0\,,\ x \in \mathfrak{X}. \end{cases}$$

が成立するとき、次の (a) , (b) が成り立つ。

(a) $$V_{\bar{\pi}}(t,x) \leq \inf_{\pi \in D_A} V_\pi(t,x) + \alpha\varepsilon \sup_{\pi \in D_A} \int_0^\infty e^{-\alpha\tau} T_\tau^\pi \rho(\bar{\pi}(\cdot,\cdot), \pi(\cdot,\cdot))(t,x) d\tau$$

(b) loss function を

$$q(t,x,a) = r(t,x,a) + \alpha\varepsilon\rho(a, \bar{\pi}(t,x)) \qquad t \geq 0\,,\ x \in \mathfrak{X}\,,\ a \in \mathcal{A}$$

と修正し、この loss function q に対応する $\pi \in D_A$ の total expected discounted loss を W_π とするとき、$\bar{\pi} \in D_A$ は修正された動的決定問題において optimal となる。つまり、

$$W_{\bar{\pi}}(t,x) = \inf_{\pi \in D_A} W_\pi(t,x) \tag{3.13}$$

が成り立つ。

3.3 一般の ε-optimal policy について

条件 A , B(a) , B(b) を仮定する。

条件 G $\exists V \in \mathcal{D}(A) \quad s.t. \quad \alpha V(t,x) = \inf_{\pi \in D_A} \{r_\pi(t,x) + A_\pi V(t,x)\} \quad , t \geq 0\,,\ x \in \mathfrak{X}$

条件 H 各 $t \geq 0, x \in \mathfrak{X}$ と条件 G の $V \in \mathcal{D}(A)$ に対して、$G.(t,x) : \mathcal{A} \longrightarrow R$ を

$$G_a(t,x) = r(t,x,a) + A_a V(t,x) \qquad a \in \mathcal{A}$$

によって定義するとき、$G.(t,x)$ は、$\mathcal{A}$ 上、下に有界かつ下半連続である。

条件 G,H を仮定するとき、Ekeland の定理により

$\forall \varepsilon > 0, \exists a^*(t,x) \in \mathcal{A} \quad s.t.$

(3.14) $a \neq a^*(t,x)$ である全ての $a \in \mathcal{A}$ に対して、

$$r(t,x,a) + A_a V(t,x) > r(t,x,a^*(t,x)) + A_{a^*(t,x)} V(t,x) - \alpha\varepsilon\rho(a^*(t,x), a)$$

条件 I $\exists \pi^* \in D_A \quad s.t. \quad \pi^*(t,x) = a^*(t,x)$

条件 J $\sup_{\pi_1, \pi_2 \in D_A} \|\rho(\pi_1(\cdot,\cdot), \pi_2(\cdot,\cdot))\| < \infty$

これらの新たな付加条件を課すことにより、次の意味での ε-optimal policy が存在する。

定理 3 条件 A , B(a) , B(b) , G , H , I , J を仮定するとき、

$\forall \varepsilon > 0$, $\exists \pi^* \in D_A$ $s.t.$

$$V_*(t,x) \geq V_{\pi^*}(t,x) - \alpha\varepsilon \int_0^\infty e^{-\alpha\tau} T_\tau^{\pi^*} \sup_{\pi \in D_A} \rho(\pi^*(\cdot,\cdot), \pi(\cdot,\cdot))(t,x) d\tau \tag{3.15}$$

$t \geq 0$, $x \in \mathfrak{X}$

証明 (3.14) と 条件 I により

$\forall a \in \mathcal{A}$, $\quad r(t,x,a^*(t,x)) + A_{a^*(t,x)} V(t,x) \leq r(t,x,a) + A_a V(t,x) + \alpha\varepsilon\rho(a^*(t,x), a)$

$\therefore$ $\forall a \in \mathcal{A}$, $\quad r_{\pi^*}(t,x) + A_{\pi^*} V(t,x) \leq r(t,x,a) + A_a V(t,x) + \alpha\varepsilon\rho(\pi^*(t,x), a)$

よって、条件 G により、次の不等式が導かれる。

$$r_{\pi^*}(t,x) + A_{\pi^*} V(t,x) \leq \alpha V(t,x) + \alpha\varepsilon \sup_{\pi \in D_A} \rho(\pi^*(t,x), \pi(t,x))$$

$$\therefore \qquad (\alpha I - A_{\pi^*}) V \geq r_{\pi^*} - \alpha\varepsilon \sup_{\pi \in D_A} \rho(\pi^*(\cdot,\cdot), \pi(\cdot,\cdot))$$

operator $(\alpha I - A_{\pi^*})^{-1}$ の単調性より、

$$\begin{aligned} V(t,x) &\geq V_{\pi^*}(t,x) - \alpha\varepsilon \int_0^\infty e^{-\alpha\tau} T_\tau^{\pi^*} \sup_{\pi \in D_A} \rho(\pi^*(\cdot,\cdot), \pi(\cdot,\cdot))(t,x) d\tau \\ &\geq V_*(t,x) - \alpha\varepsilon \int_0^\infty e^{-\alpha\tau} T_\tau^{\pi^*} \sup_{\pi \in D_A} \rho(\pi^*(\cdot,\cdot), \pi(\cdot,\cdot))(t,x) d\tau \end{aligned} \tag{3.16}$$

$\varepsilon > 0$ は任意に小さくとれるから、

$$V(t,x) \geq V_*(t,x) \tag{3.17}$$

一方、条件 G より、$\alpha V \leq r_\pi + A_\pi V$, $\forall \pi \in D_A$. [1.lemma 4.1] を用いると、$V \leq V_\pi$, $\forall \pi \in D_A$ すなわち、$V \leq V_*$ が得られる。よって (3.17) より、$V = V_*$ が成り立ち、これと (3.16) から (3.15) が導かれる。 ■

参考文献

[1] Doshi,B.T.(1976) Continuous time control of Markov processes on an arbitrary state space : discounted rewards. Ann.Statist.4:1219-1235

[2] Dynkin,E.B.(1965) Markov processes-1. Springer-Verlag,Berlin

[3] Lai,H.C.&Tanaka,K.(1991) On Continuous-Time Discounted Stochastic Dynamic Programming. Appl.Math.Optim.23:155-169